이미지 기반 영상 요약

한승윤 O, 김세연, 박선우, 이정민, 이창희
한국외국어대학교 컴퓨터공학부
hsy0320@hufs.ac.kr, ssen98123@gmail.com, 201901496@hufs.ac.kr, 202102667@hufs.ac.kr, chlee0811@hufs.ac.kr

요 약

이 연구는 음성이 없는 영상을 요약하는 새로운 방법론을 제안한다. 현대의 영상 요약 기술은 주로 음성이나 텍스트 정보에 의존하여 작동하고 있으나, 음성이 부재한 영상에는 적용하기 어렵다. 본 연구는 시각적 요소만을 활용하여 영상을 요약하는 방법론을 제시하고자 한다.

1. Introduction

현대의 영상 요약 기술은 대부분 자막이나 음성 데이터에 의존하고 있다. 이러한 방법은 음성이나 텍스트 정보가 부재한 영상에는 적용하기 어렵다. 아래 그림 1 은 음성이 있는 영상에 대해 제공되는 요약으로, 음성 및 자막을 바탕으로 요약이 제공되고 있음을 알 수 있다. 하지만, 그림 2 의 경우 음성이 없는 무성 애니메이션이기 때문에, 요약이 진행되지 않음을 확인할 수 있다. 이처럼, 무성 영화, 무성 애니메이션, 또는 CCTV 와 같은 감시 영상들은 오직 시각적 내용에만 의존해야 한다. 이러한 영상들은 기존의 자막이나 음성 기반 요약 기술로는 해결할 수 없다.

그림 1. 음성이 있는 영상에 대한 요약

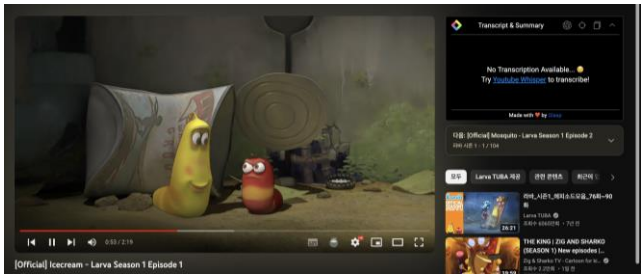

그림 2. 음성이 없는 영상에 대한 요약

따라서, 시각적 요소만을 활용하는 새로운 영상 요약 방법론의 개발은 미디어 분석, 감시 시스템, 엔터테인먼트 산업 등 다양한 분야에서 중요한 역할을 할 수 있다.

본 연구의 목적은 인간의 시각 처리 방식을 모방하여, 음성이나 텍스트 정보 없이도 영상의 내용을 효과적으로 요약하고 이해할 수 있는 새로운 영상 요약 기술을 개발하는 것이다. 인간의 뇌는 영상을 고속의 연속적인 이미지 시퀀스로 처리한다. 각각의 이미지는 독립적으로 분석되며, 이전 이미지의 정보와 연결되어 전체적인 이해를 도출한다. 이러한 인간의 시각 처리 메커니즘에 착안하여, 개별 프레임을 독립적으로 분석하고, 이를 종합적으로 요약하는 방법론을 개발하고자 한다.

인간의 시각 시스템은 원활한 영상 해석을 위해 약 2-3 프레임/초의 속도로 이미지를 처리하며, 특정 객체나 움직임에 집중할 수 있도록 해준다. 또한, 시각적 연속성을 통해 개별적인 이미지들을 하나의 연속적인 스토리로 해석하며, 시각적 기억과 예측을 통해 이미지를 인식하고 저장한 다음, 이전의 시각적 경험을 바탕으로 다가올 시각적 정보를 예측한다. 마지막으로, 우리의 뇌는 시각적 정보를 해석하여 의미를 부여하며, 이는 단순히 이미지를 '보는' 것 이상의 과정을 포함한다.

이 연구는 이미지를 인식하고, 인식한 이미지로부터 텍스트를 추출하는 BLIP 모델을 사용하여 각 프레임을 분석하고, 이미지 프레임마다 추출된 문장을 BART 를 사용하여 종합적으로 요약한다. 이러한 접근 방식은 인간의 시각 시스템을 모방하여, 영상을 특정 기준의 프레임으로 분할하고, 각각의 이미지에 대한 정보를 해석하는 방식이다. 이는 기존의 음성이나 텍스트 정보에 의존하지 않는 새로운 차원의 영상 해석을 가능하게 한다.

본 연구는 인간의 인지 과정에 대한 깊은 이해를 제공할 수 있으며, 이는 컴퓨터 비전과 인공지능 분야에 있어서 중요한 발전을 의미한다.

2. Method

2.1 이론적 배경

2.1.1 BLIP

Image to Text 를 할 때에는 BLIP(Bootstrapping Language-Image Pre-training)은 비전-언어 이해 및 생성 작업 양쪽으로 유연하게 전이되는 VLP(Vision-Language Pre-training) 프레임워크로, 웹 데이터를 효과적으로 활용하여 캡션을 부트스트랩하고, 캡션 생성자가 합성 캡션을 생성하고 필터가 노이즈가 있는 캡션을 제거한다. BLIP 은 다양한 비전-언어 작업에서 우수한 결과를 보여주고 있으며, 이미지-텍스트 검색, 이미지 캡션 생성, VQA 등의 작업에서 다른 모델들을 능가하는 성과를 보여주고 있다. 또한 BLIP 은 비디오-언어 작업으로 직접 전이되어 제로샷 방식으로 강력한 일반화 능력을 나타낸다.

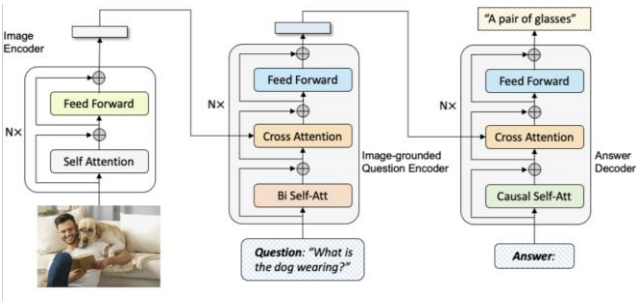

그림 3. BLIP 모델 구조

2.1.2 BART

텍스트 요약을 할 때에는 BART(Bidirectional and Auto-Regressive Transformers)를 사용하였는데, BART 의 사전 훈련은 텍스트 문서를 손상시키고 디코더의 출력과 원본 문서 간의 교차 엔트로피를 최소화하는 재구성 손실을 최적화하는 것으로 이루어진다. 이러한 방식으로 BART 는 어떤 유형의 문서 손상에도 사용될 수 있다.

BART 의 아키텍처는 Vaswani et ai. (2017)의 Transformer 디자인을 기반으로 하되, ReLU 활성화 함수를 GeLUs 로 바꾸고 매개 변수를 (0, 0.02)로 초기화한것을 제외하면 일반적인 시퀀스-투-시퀀스 Transformer 디자인을 따르고 있다.

BART 는 세분화된 텍스트 생성을 위해 미세 조정이 가능한데, 시퀀스 분류 작업, 토큰 분류 작업, 시퀀스 생성 작업, 기계 번역의 용도로 쓰일 수 있다.

2.2 데이터 전처리

이 연구에서는 '라바' 시즌 1 의 영상 데이터를 수집하여 사용하였는데, 데이터 수집은 다음과 같은 과정을 거쳤다. 먼저, openCV 라이브러리를 활용하여 영상에서 초당 2-3 프레임으로 캡처하였다. 캡처된 프레임에 대해서는 연구 목적에 맞게 엑셀 파일에 모든 장면에 대한 자세한 설명을 작성하였다. 캡처된 이미지의 각 장면에 대한 설명은 해당 장면의 주요 특징과 상황을 담고 있으며, 물리적 환경, 캐릭터들의 행동, 그리고 각 장면의 중요한 사건들에 대한 정보를 상세히 기술하였다. 이러한 설명은 각 프레임이 나타내는 시각적인 내용을 정확하게 이해하고 해석하는 데에 도움이 되도록 구성되었다. 데이터 수집의 주요 목적은 시각적인 정보를 포함한 영상의 흐름을 잘 파악하고, 해당 정보를 기반으로 이미지를 텍스트로 변환하고 이를 효과적으로 요약하는 모델을 개발하는 데에 있었다. 이를 위해 라바 시즌 1 의 다양한 장면을 포함한 데이터셋을 구축하였으며, 이는 모델의 학습과 성능 평가에 활용되었다.

3. Results

3.1 Image to Text

Image to Text 모델인 BLIP 으로 이미지 텍스트 생성 실험을 수행하였다. 아래 그림 1 은 사용하여 '라바' 애니메이션의 두 주요 캐릭터인 옐로우와 레드가 빨대 양쪽을 물고 있는 장면으로, 해당 이미지를 사용하여 두가지 모델에 대한 결과를 비교 분석하였다.

그림 4. Image to Text 모델에 사용한 이미지

표 1. Image to Text 모델 비교

모델	텍스트
BLIP-image-captioning	a cartoon character is playing with a tooth
Fine tuning + BLIP-image-captioning	red is trying to eat the remaining drink on the straw, and yellow is also trying to eat the remaining drink on the straw.

첫 번째 실험에서는 COCO 데이터셋으로만 학습

된 기본 BLIP-Image-Captioning 모델을 사용했다. COCO 데이터셋은 다양한 일상적인 장면과 객체를 포함하고 있어, 일반적인 이미지 인식 및 텍스트 생성에 효과적이다. 이 설정에서 모델은 'a cartoon character is playing with a tooth'라는 텍스트를 생성했다. 이 텍스트는 이미지의 기본적인 내용을 어느 정도 포착했으나, 라바 캐릭터의 구체적인 특징과 상황을 정확히 반영하지는 못했다.

두 번째 실험에서는 COCO 데이터셋으로 학습된 BLIP 모델에 라바 캐릭터에 대한 사전 학습을 추가적으로 적용했다. 중요한 점은 모델의 overfitting 을 방지하기 위해 일부 layer 들을 freezing 상태로 두고 학습을 진행했다는 것이다. 이러한 방법은 모델이 기존에 학습한 일반적인 지식을 유지하면서 새로운 정보를 효율적으로 통합할 수 있도록 한다. 이 실험에서는 모델이 'red is trying to eat the remaining drink on the straw, and yellow is also trying to eat the remaining drink on the straw.'라는 훨씬 더 상세하고 정확한 텍스트를 생성했다.

두 실험 결과를 비교해보면, Pretrained BLIP 모델이 더 정확하고 상세한 정보를 포함한 캡션을 생성했다는 것을 알 수 있다. 기본 BLIP 모델은 이미지의 일반적인 상황은 인식했지만, 라바 캐릭터의 구체적인 특성과 상황에 대한 이해는 부족했다. 반면, Pretrained 모델은 라바 캐릭터의 구체적인 특징과 행동을 정확히 인식하고 이를 캡션에 반영했다. 이러한 결과는 특정 캐릭터와 상황에 대한 모델의 이해력을 향상시키는 데 사전 학습이 필요함을 의미한다. 또한, 모델의 8 개 layer (encoder.layer.0 부터 encoder.layer.3 까지, bert.encoder.layer.0 부터 bert.encoder.layer.3 까지)를 freezing 하여, 모델이 기존에 학습한 일반적인 파라미터를 유지하면서 새로운 정보를 효과적으로 통합할 수 있도록 하였다. 즉, 적절한 layer freezing 기법이 중요한 역할을 한다는 것을 보여준다.

이러한 발견은 BLIP 모델의 활용성과 유연성을 입증하며, 특정 도메인에 대한 사전 학습과 적절한 layer freezing 기법의 적용이 모델의 성능을 향상시킬 수 있음을 시사한다.

3.2 Text Summarization

Text Summarization 에는 두 가지 모델을 비교하였다. 먼저, CNN 뉴스 데이터로 사전 학습된 BART 와 T5 모델을 사용하여 성능을 비교하였다.

모델의 input 으로 CNN 뉴스 기사의 일부분을 사용하였고, 아래 표와 같은 요약이 제공되었다.

표 2. Text Summarization 모델 비교 (CNN)

모델	텍스트
BART	Liana Barrientos, 39, is charged with two counts of "offering a false instrument for filing in the first degree" In total, she has been married 10 times, with nine of her marriages occurring between 1999 and 2002. She is believed to still be married to four men.
T5	A woman has pleaded not guilty to married 10 times in a row over immigration.

T5 모델과 비교했을 때 BART 모델의 결과는 기사의 중요한 요소들을 포함하고 있었으며, 특히 법적 과정과 결혼의 횟수 및 기간에 대한 정보를 효과적으로 요약하였다. 이에 반해, T5 모델의 요약은 상대적으로 덜 구체적이며, 핵심적인 법적 측면과 결혼 횟수에 대한 정보가 누락되었다. 이러한 결과를 바탕으로, BART 모델이 이 경우 더 정확하고 상세한 요약을 제공하는 것으로 평가하였다.

두번째로는 '라바' 애니메이션의 Image to Text 변환을 통해 얻은 데이터를 요약하는 성능을 비교하였다.

모델의 input 은 레드와 옐로우가 스파게티 면을 먹는 장면에 대한 텍스트이다.

표 3. Text Summarization 모델 비교 (라바)

모델	텍스트
BART	Red is biting into a spaghetti noodle that has fallen from the sky, and yellow is looking at it. Yellow is looking up with an ecstatic expression on his face, while red is lying on the ground. red and yellow are tied to the spaghetti noodles, and yellow has fallen apart.
T5	Red is biting into a spaghetti noodle that has fallen from the sky . Yellow is looking up with an ecstatic expression on his face, while red is lying on the ground . Red and yellow are tied to the spaghetti noodles.

두 모델의 결과를 비교했을 때, BART 모델의 결과는 원본 텍스트의 주요 내용을 잘 포착하고 있으며, 두 캐릭터의 행동과 상태를 명확하게 요약하고 있음을 확인할 수 있다. T5 모델은 비슷한 내용을 담고 있지만, 상세한 부분에서는 BART 모델만큼 구체적이지 않은 것을 확인할 수 있다.

이러한 비교를 통해 BART 는 텍스트의 복잡한 구조와 다양한 정보를 효율적으로 처리하고 요약하는 능력을 가지고 있음을 확인할 수 있었다. 특히, 이미지에서 추출된 텍스트 데이터에서 세부 사항과 정황을 정확하게 포착하고 표현하는 능력이 BART 모델의 장점임을 확인하였다. 즉, BART 모델이 복잡하고 다양한 정보가 포함된 텍스트에서 핵심적인 요소들을 포착하고 효과적으로 요약할 수 있음을 시사한다.

4. Discussion

본 연구는 제한된 데이터셋으로 인해 충분한 학

습을 수행하지 못한 점에서 한계를 가지고 있다. 이는 연구의 결과를 최대한 활용하는 데 있어 제약으로 작용했다. 그럼에도 불구하고, 데이터셋의 다양화와 확장이 이루어진다면, 이 연구는 훨씬 더 깊이 있는 분석과 더욱 정교한 결과를 제공할 수 있는 잠재력을 가지고 있다고 볼 수 있다. 특히, 이 연구가 음성이 없는 영상에 대한 새로운 접근 방법을 제시하고 있다는 점에서 큰 의의를 찾을 수 있다. 인간의 시각 처리 원리에 기반한 이 방법론은 무성영화, 애니메이션, 감시 영상 등 다양한 형태의 음성이 없는 미디어에 효과적으로 적용될 수 있으며, 이는 기존의 연구들에서는 다루지 않았던 새로운 영역을 개척하는 것을 의미한다.

더 나아가, 본 연구의 결과는 영상 콘텐츠의 해석과 정보 제공 방식에 새로운 차원을 제시한다. 특히, 넷플릭스와 같은 스트리밍 서비스에서 영상의 효과음을 자막으로 정확하게 표현하는 데에 중요한 역할을 할 수 있다. 영상의 시각적 요소를 정밀하게 해석하고 이를 텍스트로 변환하는 기술은, 효과음이나 배경 소리를 더욱 정확하고 생생하게 전달하는 데 기여할 것이다. 또한, 이 연구는 기존의 영상 요약 기술과 결합될 경우, 더욱 풍부하고 다층적인 정보 제공이 가능해질 것이며, 이는 광고, 교육, 엔터테인먼트 등 다양한 분야에서의 응용 가능성을 열어준다.

이러한 관점에서 볼 때, 본 연구는 음성이 없는 영상을 해석하고 요약하는 분야에서 새로운 방법을 제공했다고 할 수 있다. 더 많은 데이터와 더욱 발전된 학습 방법론을 통해, 이 연구는 향후 더 큰 잠재력을 발휘하며, 음성이 없는 미디어에 대한 이해와 활용 방식을 혁신적으로 변화시킬 수 있을 것으로 기대된다.

감사의 글

이 연구가 완성될 수 있도록 함께 노력해주신 저희 팀원들에게 깊은 감사를 드립니다. 이 연구가 음성이 없는 영상을 요약하는 모델 개발에 기여할 수 있길 바랍니다.

참고문헌

[1] Junnan Li, Dongxu Li, Caiming Xiong, Steven C. H. Hoi. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." arXiv preprint arXiv:2201.12086 (2022). DOI: 10.48550/arXiv.2201.12086

[2] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL].

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG].

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv:1810.04805 [cs.CL].

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). "Attention Is All You Need." In Advances in Neural Information Processing Systems (pp. 5998-6008).

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).